2017-02

Making sense of words: a robotic model for language abstraction

Stramandinoli, F

10.1007/s10514-016-9587-8
Autonomous Robots
Springer Science and Business Media LLC

Auton Robot (2017) 41:367–383
DOI 10.1007/s10514-016-9587-8

Making sense of words: a robotic model for language abstraction

Francesca Stramandinoli[1,2] · **Davide Marocco**[2] · **Angelo Cangelosi**[2]

Received: 31 August 2015 / Accepted: 23 June 2016 / Published online: 1 July 2016

Abstract Building robots capable of acting independently in unstructured environments is still a challenging task for roboticists. The capability to comprehend and produce language in a 'human-like' manner represents a powerful tool for the autonomous interaction of robots with human beings, for better understanding situations and exchanging information during the execution of tasks that require cooperation. In this work, we present a robotic model for grounding abstract action words (i.e. USE, MAKE) through the hierarchical organization of terms directly linked to perceptual and motor skills of a humanoid robot. Experimental results have shown that the robot, in response to linguistic commands, is capable of performing the appropriate behaviors on objects. Results obtained in case of inconsistency between the perceptual and linguistic inputs have shown that the robot executes the actions elicited by the seen object.

Keywords Developmental robotics · Language modeling · Sensorimotor knowledge · Symbol grounding · Embodiment

✉ Francesca Stramandinoli
francesca.stramandinoli@iit.it

Davide Marocco
davide.marocco@plymouth.ac.uk

Angelo Cangelosi
A.Cangelosi@plymouth.ac.uk

[1] iCub Facility Department, Istituto Italiano di Tecnologia, 16163 Genoa, Italy

[2] Centre for Robotics and Neural Systems, Plymouth University, Devon PL48AA, UK

1 Introduction

The decreasing costs of sensor technology and computational power achieved during the last decade is leading to a new generation of robots that can act and perform behaviors independently. For the autonomous interaction of a robot with humans, a combination of verbal and non-verbal communication skills is needed. Indeed, robots that will help humans in everyday life need to be able to communicate appropriately. Personal domestic robots endowed with the capability to comprehend and produce language in a 'human-like' manner can facilitate the interaction with human beings. However, the implementation of behaviors that can make the interaction with robots natural and intuitive for their human users, is one of the challenges that roboticists are still facing.

Different directions have been taken in the attempt to model language in artificial systems. Pure symbolic approaches (Landauer and Dumais 1997), by studying language in isolation from other cognitive skills, as mere symbol manipulation capabilities, have failed in their implementation in robots. According to the embodied approach instead, language has to be grounded in perception and motor knowledge (Barsalou 1999). However, the representation of abstract concepts poses a classical challenge for grounded theories of cognition. Indeed, given their weak perceptual and cognitive constraints with the physical world, abstract concept acquisition cannot be simply resolved by directly linking words to the entities and concepts to which they refer.

We present a model for grounding abstract action words (i.e. USE, MAKE) in sensorimotor experience. In particular, we propose a general mechanism for grounding abstract action words through the hierarchical organization of terms directly linked to the perceptual and motor knowledge of a humanoid robot (Cangelosi et al. 2010).

The outline of this paper is as follows. In Sect. 2 studies on the embodiment and combinatoriality of language and the motor system are presented; the section also describes the goal of the study. In Sect. 3 we present related computational models, while Sect. 4 describes the model we propose for the grounding of abstract action words. In Sect. 5 we introduce the training of the model. Section 6 contains the results of the study, while in Sect. 7 we draw conclusions and present an outlook on future work.

2 Embodiment and combinatoriality of language and the motor system

Studies presented in neuroscience (Pulvermüller et al. 2001; Hauk et al. 2004; Tettamanti et al. 2005; Buccino et al. 2005) and the behavioral sciences (Buccino et al. 2005; Scorolli and Borghi 2007) have demonstrated that language is embodied in perceptual and motor knowledge. According to this embodied perspective, language skills develop together with other cognitive capabilities and through the sensorimotor interaction of an agent with the environment. In such a context, particular attention has been given to the representation of action words, which are verbs referring to actions like pick, kick, lick. Through electroencephalography (EEG) recordings it has been shown that the processing of action words causes differential activation along the motor strip in the brain, with strongest in-going activity occurring close to the cortical representation of the body parts (e.g. hands, legs, lips) primarily used for carrying out the actions described by the processed verbs (Pulvermüller et al. 2001). Other studies have shown that action word meanings have correlates in the somatotopic activation of the motor and premotor cortex (Hauk et al. 2004). Moreover, transcranial magnetic stimulation (TMS) studies and behavioral experiments have shown that the processing of action-related sentences modulates the activity of the motor system (Buccino et al. 2005); according to the effector used in the action described by the processed action word, different sectors of the motor system are activated (Buccino et al. 2005).

Psychological studies and theories on the embodiment of language have been proposed as well. According to the perceptual symbol systems (PSSs) theory, conceptualization requires the simulation of past experience (Barsalou 1999). For example, when thinking about an object, the neural patterns in the brain formed during earlier experience done with it, are reactivated. The neural underpinnings of this simulation could be found in wide neural circuits that involve canonical and mirror neurons (Rizzolatti et al. 1996). In other studies performed in the field of language comprehension (Glenberg and Kaschak 2002), it has been observed that sentences are understood by creating a simulation of the actions that underlie them (Action-sentence Compatibility Effect).

In contrast to other forms of communication, language is a combinatorial system that permits the conveyance of new messages and concepts by combining words together. A finite number of terms (i.e. lexicon) can be combined and permuted according to specific structural rules (i.e. grammar) in order to convey new meanings (Pinker 1994). Growing evidence has suggested that the human motor system is also hierarchically organized; that is, low level motor primitives can be integrated and recombined in different action sequences in order to perform novel tasks (Mussa-Ivaldi and Bizzi 2000). Studies investigating how the brain accomplishes action organization have been proposed in Grafton and Hamilton (2007). The authors have argued that action organization is based on a hierarchical model, which includes different levels of motor control: (i) the level of action intention, (ii) the level of object-goal to realize the intention, (iii) the level of kinematic that represents the actions required to achieve the movement goal, and (iv) the level of muscle that coordinates the activation of muscles to produce the movement goal. Moreover, in DeWolf and Eliasmith (2011) authors have presented the Neural Optimal Control Hierarchy (NOCH), proposed as a framework for biologically plausible models of neural motor control. The simulation of the NOCH framework has suggested that the integration of control theory with the basic anatomical elements and functions of the motor system can be useful to have a unified account on a variety of motor system data. In our work for the implementation of the motor behaviors performed by the robot we were inspired by the 'schema theory' proposed in Arbib and Érdi (1998), according to which complex human behavior are built through the hierarchical organization of the motor system within which reusable motor primitives can be re-organized into different motor sequences. For example, when we want to drink a cup of coffee we segment this complex action into a combination of low level primitives, like for example the reaching, grasping and bringing to the mouth of the cup. This theory has inspired many other studies on the hierarchical organization of the motor system. For example, in Mussa-Ivaldi and Bizzi (2000) it has been suggested that low level motor primitives can be integrated and recombined in different action sequences in order to perform novel tasks. The authors have proposed that modular primitives are combined in the spinal cord in order to build the internal representation of a limb movement.

Taken together these studies suggest that both language and the biological motor system are based on hierarchical recursive structures that can enable the grounding of concepts and language in perception and motor knowledge (Cangelosi et al. 2010).

2.1 Embodied abstract language and hierarchical categories

The representation of abstract concepts poses a challenge for grounded theories of cognition. Different scholars have claimed that embodiment plays an important role even in representing abstract concepts; theories based on "simulations" (Barsalou 1999), "metaphors" (Lakoff and Johnson 1980) and "actions" (Glenberg and Kaschak 2002) have been presented. In Barsalou (1999) it has been proposed that some abstract concepts arise from simulation processes of internal and external states. In particular, abstract concepts require to capture complex multi-modal simulations of temporally extended events, with simulations of introspections being central (Barsalou 1999); indeed, introspection gives access to subjective experiences linked to abstract concepts (Wiemer-Hastings et al. 2001). Considering that abstract concepts contain more information about introspection and events, simulators for abstract words develop to represent categories of internal experience (Barsalou 2009). Hence, according to this approach, abstract concepts, differently from concrete ones, require the activation of situations and introspections. Another theory proposed on the embodiment of abstract language revolves around the concept of "metaphor". According to this approach, there are image-schemas derived from sensorimotor experience that can be transferred to experience which is not truly sensorimotor in nature (Lakoff and Johnson 1980). Human beings have an extensive knowledge about their bodies (e.g. eating) and situations (e.g. verticality) that they can use to metaphorically ground abstract concepts (Barsalou 2008); for example, *love* can be understood as eating (e.g. "being consumed by a lover"), while an affective experience like *happy*/*sad* can be understood as verticality (e.g. "up/down"). The idea that embodiment plays an important role for representing abstract concepts has been supported by other scholars. For example according to Glenberg and Kaschak (2002), sentences including both concrete and abstract words are understood by creating a simulation of the actions that underlie them. Indeed, abstract concepts containing motor information can be represented by using modal symbols. Moreover, through behavioral and neurophysiological studies it has been shown that the comprehension of abstract words activates the motor system (Glenberg et al. 2008). Hence, according to these studies, abstract concepts, similarly to concrete ones, can be grounded in perception and action.

However, other scholars have suggested that abstract concepts are only partially grounded in sensorimotor experience. Indeed, according to the theory proposed in Dove (2011), although most concepts require two types of semantic representations [i.e. (i) based on perception and motor knowledge, and (ii) based on language], abstract concepts tend to depend more on linguistic representations. According to the Language and Situated Simulation (LASS) theory presented in Barsalou et al. (2008), both the sensorimotor and linguistic systems are activated during language processing. However, concrete and abstract concepts activate different brain areas depending on their contents; moreover, according to the task to be performed (e.g. lexical decision vs. imagination task) there is a higher engagement of linguistic versus sensorimotor areas. For example, in lexical decision tasks using the linguistic system represents a shortcut as it allows to respond immediately without necessarily accessing the sensorimotor information used for conceptual meaning representation (Borghi et al. 2014). Other scholars have proposed the "Words As social Tools" (WAT) theory (Borghi and Binkofski 2014) that accounts how different kinds of abstract concepts and words (ACWs) are represented; words represent tools that permit to act in the social world. Indeed, the acquisition of ACWs relies more on language and on the contribution that other people can provide to the clarification of word meanings. In Kousta et al. (2011) authors have claimed that words which refer to emotions should be categorized in a group distinct from concrete and abstract words. This proposal was motivated by the fact that concrete, abstract and emotion words received different ratings in term of concreteness, imageability and context availability.

Given the current debate in the field and the complexity of the matter, the representation of abstract concepts is increasingly proving to be an extremely complex task. Studies conducted on children's early vocabulary acquisition (McGhee-Bidlack 1991) have shown that, when children learn to speak, they first learn concrete nouns (e.g. object's name) and then the abstract ones (e.g. verbs). While concrete terms refer to tangible entities characterized by a direct mapping to perceptual-cognitive information, abstract words referring to many events, situations and bodily states (Barsalou 1999; Wiemer-Hastings and Xu 2005) have weaker perceptual-cognitive constraints with the physical world. Hence, during the process of word meaning acquisition, the mapping of perceptual-cognitive information related to concrete concepts into the linguistic domain occurs earlier than the mapping of perceptual-cognitive information related to abstract concepts. However, the transition from highly concrete concepts to the abstract ones is gradual; that is, the categorization of concrete and abstract terms cannot be simply regarded as a dichotomy (Wiemer-Hastings et al. 2001) but there is instead a continuum in the level of abstractness, according to which all words can be categorized. The most influential theories proposed on the learning and representation of categories/concepts are the Prototype Theory and the Exemplar Theory. According to the Prototype Theory, concepts are represented by characteristic features, which are weighted in the definition of prototypes used for judging the membership of other items to the same category (Rosch and Mervis 1975). According to the Exemplar Theory, a

concept is represented by the exemplars of the categories (i.e. a set of instances of it) stored in the memory. A new item is classified as a member of a category if it is sufficiently similar to one of the stored exemplars in that category (Nosofsky et al. 1992). In the context of the Exemplar Theory, it has been proposed the instantiation principle (Heit and Barsalou 1996), according to which the representation of superordinate concepts evoke detailed information about its subordinate members (i.e. exemplars). In Murphy and Wisniewski (1989) the authors conducted a categorization study that has shown that when an object is placed in an inappropriate scene, there is more interference for the identification of the exemplars of superordinate concepts than for basic level concepts. According to the classical theory of categorization, words can be organized in hierarchically structured categories (Gallese and Lakoff 2005) along which the level of abstraction can vary considerably. For example, in the hierarchy of categories "furniture/chair/rocking chair", "furniture" is a superordinate word (i.e. generalization w.r.t. the concept related to the basic word "chair") while "rocking chair" is a subordinate word (i.e. specialization w.r.t. the concept related to the basic word "chair"). In this framework, basic and subordinate words (e.g. "chair", "rocking chair"), refer to single entities and they can be seen as more concrete words than the superordinate ones (e.g. "furniture") which refer to sets of entities that differ in shape and other perceptual characteristics (Borghi et al. 2011). Moreover, categories like "furniture" that do not have corresponding motor programs for interacting with them, represent general and abstract concepts.

Among the different lexical categories (i.e. noun, verb, adjective, adverb, etc.), abstract action words represent a class of terms distant from immediate perception that describe actions (i.e. verbs) with a general meaning (e.g. USE, MAKE) and which can be referred to several events and situations (Barsalou 1999; Wiemer-Hastings et al. 2001). Therefore, they cannot be directly linked to sensorimotor experience through a one-to-one mapping with their physical referents in the world. For example, the meaning of words like USE and MAKE is general and it depends on the context in which they occur (Barsalou et al. 2003). In a scenario in which a person is interacting with a set of tools, the meaning of USE is specified by the particular tool employed during the interaction (e.g. USE [a] KNIFE, USE [a] BRUSH), while the meaning of MAKE depends on the outcome of interactions (e.g. MAKE [a] SLICE, MAKE [a] HOLE).

2.2 Goal of the study

In this work we present a model based on Recurrent Neural Networks (RNN) for the grounding of abstract action words (i.e. USE and MAKE) achieved through the hierarchical organization of words directly linked to perceptual and motor knowledge of a humanoid robot; indeed, building on our previous work (Cangelosi and Riga 2006; Stramandinoli et al. 2012) we attempt to extend the "grounding transfer mechanism" from sensorimotor experience to abstract concepts. Our proposal is that words that refer to objects and actions primitives can be grounded in sensorimotor experience, while abstract action words require linguistic information as well. Linguistic information permits to create the semantic referents of terms that cannot be directly mapped into their referents in the physical world (Stramandinoli et al. 2010, 2012; Stramandinoli 2014). The semantic referents of these words are formed by recalling and reusing the motor and perceptual knowledge directly grounded during previous experience of the robot with the environment. Words directly linked to sensorimotor experience, combined in hierarchical structures through language, permit the indirect grounding of abstract action words. We propose such a hierarchical organization of concepts as a possible account for the acquisition of abstract action words in cognitive robots.

The aim of this work is twofold. On the one hand, the robotic platform is enabled to ground the meaning of abstract action words and scaffold more complex behaviors through the sensorimotor interaction with the environment; on the other hand, the proposed model permits the investigation of the relation between perceptual and motor categories, and the development of conceptual knowledge in a humanoid robot.

3 Related computational models

Recently, roboticists have started to investigate some of the issues related to language development. However, attempts to model the acquisition of abstract language in robots are in fact non-existent. Different models have focused on the acquisition of words related to objects and actions but none of them addressed the problem of grounding abstract categories. For example, Sugita and Tani (2005) have proposed a model for the acquisition of the meaning of simple linguistic commands; a mobile robot acquires the meaning of two-word sentences through the translation of linguistic commands into context-dependent behaviors. In Yamashita and Tani (2008) a humanoid robot has learned to generate object manipulation behaviors by a functional hierarchy which self-organizes through multiple time-scales in the activity of the neural network based model. In Dominey et al. (2009) a model for the learning of a cooperative assembly task has been presented; a user can guide the robot through an arbitrary, task relevant, motor sequence via spoken commands and the robot can acquire on the fly the meaning of novel linguistic instructions and new behavioral skills by grounding the new commands in combinations of pre-existing motor primitives. In Kalkan et al. (2013) the interactions of a robot with its environment have been used to create concepts typically represented by

verbs in language. In Yürüten et al. (2012) a model for the learning of adjectives and nouns from affordances has been presented; the iCub humanoid robot is enabled to learn nouns and adjectives from sensorimotor interactions and to predict the effects of the interaction with objects (e.g. labeled as verbs). All the described models have focused on the learning of different lexical categories (e.g. adjectives, nouns and verbs) which can be directly mapped into physical referents in the real world (i.e. concrete concepts). In Stramandinoli et al. (2012) sequences of linguistic inputs consisting of verbs only, have led to the grounding of "higher-order" concepts (e.g. ACCEPT, REJECT, etc.) grounded in basic motor primitives (e.g. PUSH, PULL, etc.). Indeed, in contrast to basic concepts that can be directly grounded in the sensorimotor experience of an agent, "higher-order" concepts are typically based on combinatorial aspects of language. Higher-order symbolic representations were indirectly grounded in action primitives directly grounded in sensorimotor experience. Simulation results have shown that motor primitives have different activation patterns according to the action's sequence in which they are contained. However, recently it has been argued that there is a need for modeling work in the context of cognitive robotics experiments on language learning (i.e. for the grounding of both concrete and abstract concepts) to explicitly take into account the richer human embodiment (Thill et al. 2014).

4 Model description

In this paper we present a robotic model based on Recurrent Neural Networks (Jordan 1986; Elman 1990) for the grounding of abstract action words in a humanoid robot; the grounding of abstract action words is achieved through the integration of different input signals (i.e. vision, proprioception and language). Although the proposed experimental setup is limited, given the exemplification made in the representation of the multi-modal inputs, it attempts to suggest a general mechanism for grounding abstract action words through the combination of perceptual knowledge and motor primitives. Indeed, abstract action words are grounded by linking non-verbal knowledge, both perceptual (e.g. visual features of objects like KNIFE, BRUSH, etc.) and behavioral (e.g. action primitives like CUT, PAINT, etc.), to language. We conducted our experiments using the iCub humanoid robot, an open-source platform for research in embodied cognition, artificial intelligence and brain inspired robotics research (Metta et al. 2008). The iCub software architecture is based on YARP (Metta et al. 2006) which is an open source and multi-platform framework for humanoid robotics, consisting of a set of libraries, protocols and tools that support distributed computation and that can be used for inter-process communication across a network of machines. The proposed model represents the first attempts in grounding the meaning of abstract action words in perceptual and motor experience.

4.1 Software architecture

In this work the grounding of abstract action words is achieved through the integration of the linguistic, visual and proprioceptive input modalities in a recurrent artificial neural network model. The visual and motor inputs have been recorded from the iCub sensors, while the linguistic inputs have been encoded as binary vectors for which the "one-hot" encoding has been adopted. The general overview of the implemented software architecture is presented in (Fig. 1). The iCub robot exchanges information with our software architecture through the *iCub Module* that handles the exchange of data from/to the robot. In particular, the *iCub Module* sends the proprioceptive data recorded from the robot's encoders to the *Neural Network Model*. The *Neural Network Model* computes the new joint values and the *iCub Module* sends them back to the robot. In addition to that, the *iCub Module* sends the visual stream read from the iCub cameras to the *Object Detector* module, which classifies objects according to their features and sends the generated visual input to the *Neural Network Model*. Additionally, the *Object Detector* module extracts the position of the segmented objects and sends this information to the *Head Tracker* module that moves the head of the iCub robot to the position received on-line. The visual stream read from the iCub cameras and the segmented objects are displayed through *yarpview* devices, which are the image viewers provided by the YARP middleware (Metta et al. 2006). For the implementation of the *Object Detector* and *Head Tracker*, we used some of the modules available in the iCub software repository that were adapted to be integrated with our software architecture.

4.2 Neural network model

For modeling the mechanisms underlying motor and linguistic sequences processing, Partially Recurrent Neural Networks (P-RNNs) have been used (Jordan 1986; Elman 1990). Our model is based on a three-layer Jordan P-RNN (Jordan 1986), characterized by feedback connections from the output to the input units (Fig. 2). A Jordan network is a discrete-time P-RNN in which the processing occurs in discrete steps and the relation between input/output units is governed by a functional equation that can be either linear or non-linear. The activation of the output units at time $(t-1)$ are available in the input layer (i.e. state units) at time (t) via connections which may be modified during the training of the network. The feedback of the output neurons allows the network's input units to see the previous output, and hence the

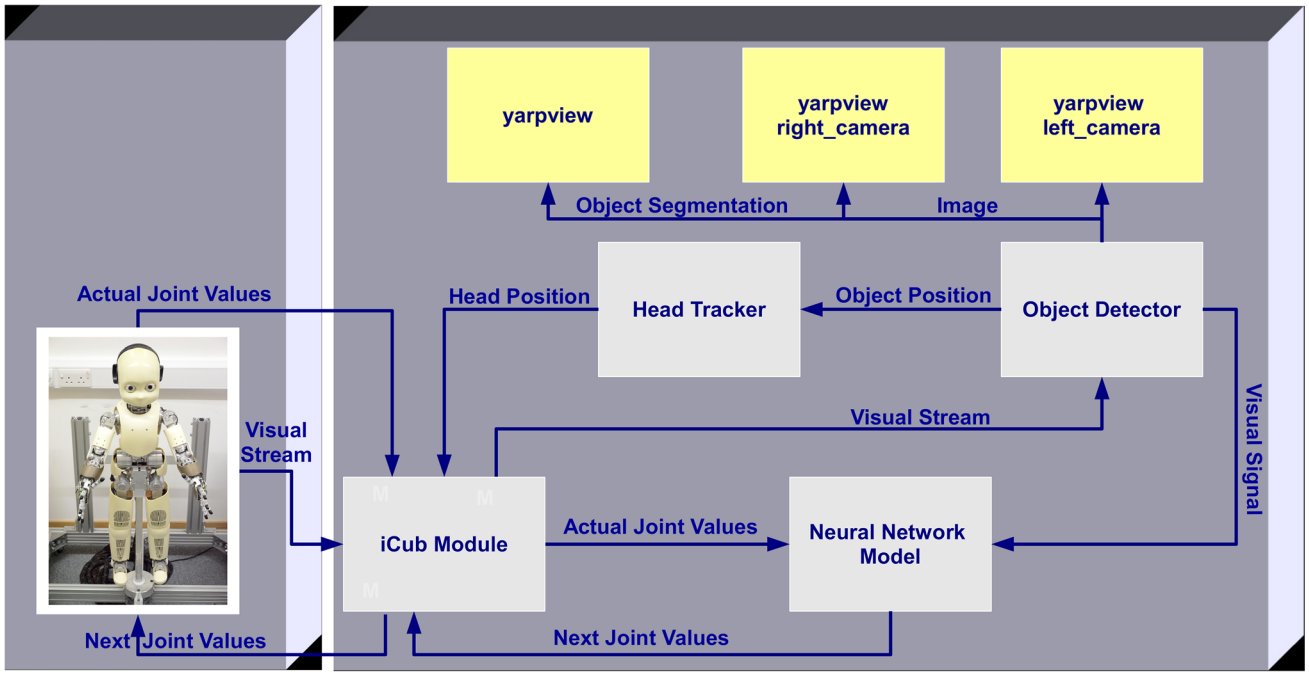

Fig. 1 Overview of the software architecture

subsequent behavior can be shaped by the previous responses of the robot.

4.2.1 Input and output coding

The input layer of the neural network consists of five units (Fig. 2) which are action's names (14 neurons), joint's angles (7 neurons), object's names (12 neurons), object's features (16 neurons) and state units (7 neurons), respectively. Further details on the different input modalities are provided below:

- *Language*: The linguistic input consists of sequences of words (i.e. verbs and nouns) arranged in two separate units of the network, which are action's names and object's names (Cangelosi and Parisi 2004). Experiments on the neural processing of verbs and nouns have shown that the left temporal neocortex plays a crucial role for nouns processing, while action's words processing involves additional regions of the left dorsolateral prefrontal cortex (Perani et al. 1999). The model was conceived with two different linguistic input units (a-priori knowledge of word's classes) in order to be able to analyze the activation values of hidden units for different classes of words.
- *Proprioception*: The proprioceptive data (i.e. joint angles of the robot's right arm) were recorded from the iCub's sensors while the robot performed target action primitives. Additional details on how we recorded the motor data are provided in Sect. 4.2.2.
- *Vision*: From the visual stream captured by the robot's cameras, object features (i.e. dimension, color and shape) were extracted. In Sect. 4.2.3 additional details on how we generated the visual input are provided.

The neural network outputs words associated to actions and objects, motor responses and the representation of object features. The proprioceptive output is sent back to the state units in the input layer by copying it; the state units contain the activation values of the proprioceptive output units of the network at time $(t-1)$ that become available to the input layer at time (t). The hidden units of the model (13 neurons), by integrating perceptual, motor and linguistic knowledge, encode the meaning of words. The selected number of hidden neurons was large enough to ensure a sufficient number of degrees of freedom for the network function and small enough to minimize the risk of loss of generalization.

4.2.2 Proprioceptive data set

For initiating the physical interaction of the robot with the environment, we have assumed that the iCub has already developed some basic skills (i.e. motor primitives like PUSH, PULL, etc.). For the performance of more complex behaviors the robot combines motor primitives into action sequences. Indeed, by exploiting the results presented in Cangelosi and Riga (2006) and Stramandinoli et al. (2012), action primitives (e.g. CUT, HIT, PAINT, etc.) are built by combining low level motor primitives (e.g. PUSH–PULL, LIFT–LOWER,

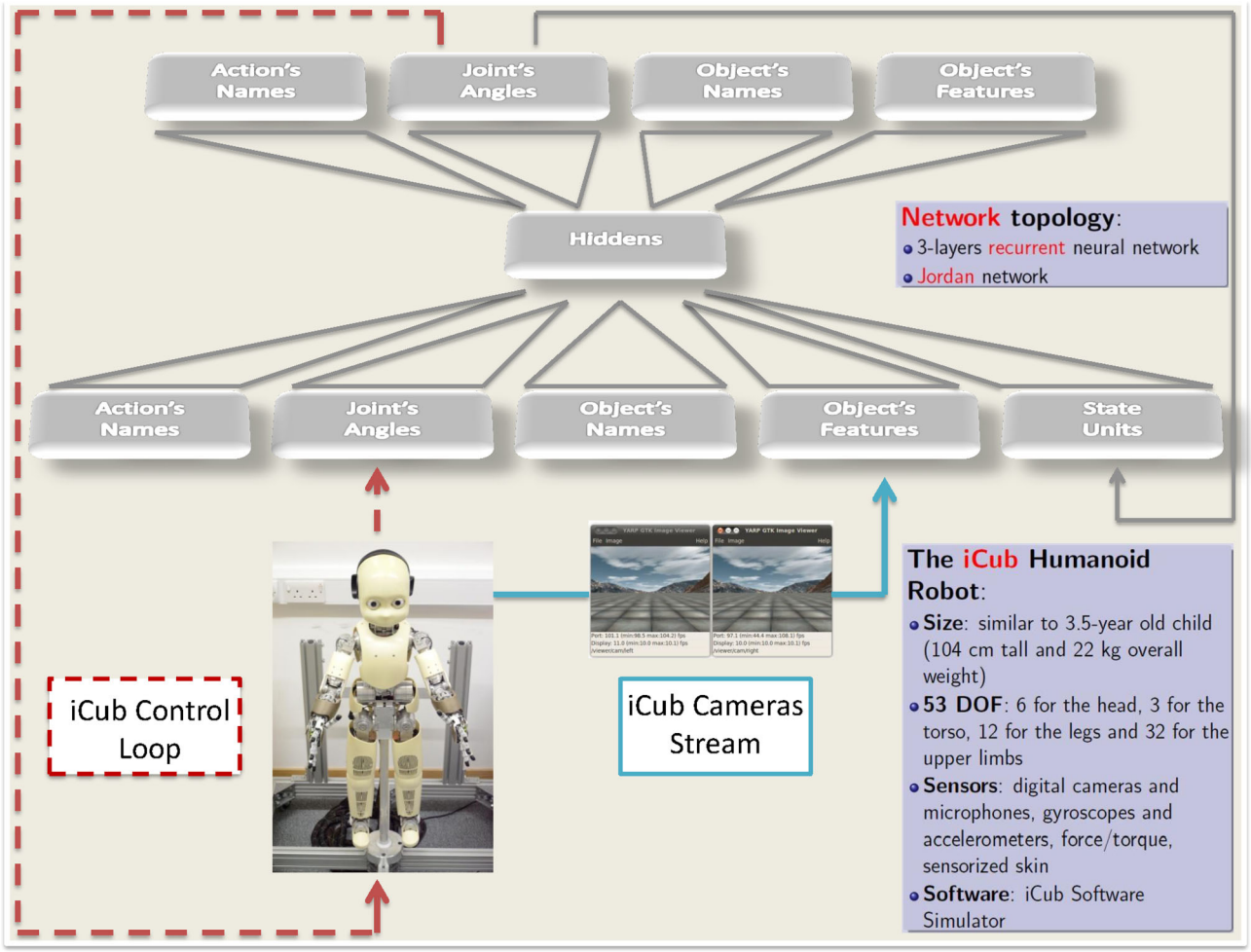

Fig. 2 Partially recurrent neural network model

MOVE_L–MOVE_R) iterated for a certain number of time steps. For example, the CUT action is built by iterating the PUSH–PULL motor primitives several times. In particular, each training sequence for the motor data consists of six elements, which corresponded to three iterations of the same action (e.g. the training sequence for the CUT action consists of PUSH–PULL, PUSH–PULL, PUSH–PULL).

Motor primitives were planned by defining the iCub end-effector pose (i.e. position x_d and orientation α_d in the 3D Cartesian space) corresponding to the movement to be performed by the arm of the robot (Oztop and Arbib 2002). Position and orientation (Eq. 1) refer to the root frame attached to the waist of the iCub; the orientation α_d, is represented in axis/angle notation [three components for the rotation axis (i.e. $\alpha_x, \alpha_y, \alpha_z$) and one component for the rotation angle expressed in radians (i.e. θ)]. The desired task-space behavior is mapped into the appropriate joint trajectories (i.e. q) by solving the inverse kinematics problem that determines the values of the seven joints (Eq. 3) of the iCub right arm (Shoulder Pitch θ_{sp}, Shoulder Roll θ_{sr}, Shoulder Yaw θ_{sy}, Elbow θ_e, Wrist Pronosupination θ_{wpr}, Wrist Pitch θ_{wp} and Wrist Yaw θ_{wy}). Given the position $x_d \in \mathbb{R}^3$ and orientation $\alpha_d \in \mathbb{R}^4$ of the iCub end-effector for different motor primitives:

$$x_d = [xyz]^T \in \mathbb{R}^3 \tag{1}$$

$$\alpha_d = ([\alpha_x \alpha_y \alpha_z]^T, \theta) \in \mathbb{R}^4 \tag{2}$$

the joint space vector $q \in \mathbb{R}^7$ is determined by solving the inverse kinematics problem by using the Cartesian interface available in the iCub software repository (Pattacini et al. 2010):

$$q = [\theta_{sp}\ \theta_{sr}\ \theta_{sy}\ \theta_e\ \theta_{wpr}\ \theta_{wp}\ \theta_{wy}]^T \in \mathbb{R}^7 \tag{3}$$

During the recording of joint values, motor primitives started and ended from the same home position (i.e. $x = -0.29$, $y = 0.16$, $z = 0.0$); the orientation of the end effector was fixed (i.e. $\alpha_x = 0.12$, $\alpha_y = 0.76$, $\alpha_z = -0.64$, $\theta = 3.0$). Half of the primitives were iterative (Table 1a) and

Table 1 Action's name, object's name and positions from which the iCub arm joint values were recorded

Action's name	Position			Object's name
	x	y	z	
(a) *Iterative actions*				
CHOP	−0.24	0.16	0.0	KNIFE
	−0.29	0.16	0.0	
CUT	−0.21	0.16	0.0	SAW
	−0.29	0.16	0.0	
HIT	−0.29	0.16	0.05	HAMMER
	−0.29	0.16	0.0	
POUND	−0.29	0.16	0.08	STONE
	−0.29	0.16	0.0	
DRAW	−0.29	0.21	0.0	PENCIL
	−0.29	0.16	0.0	
PAINT	−0.29	0.24	0.0	BRUSH
	−0.29	0.16	0.0	
(b) *Non-iterative actions*				
SLICE	−0.24	0.13	0.0	SLICER
	−0.29	0.16	0.0	
SLIT	−0.21	0.11	0.0	BLADE
	−0.29	0.16	0.0	
HOLE	−0.29	0.1	0.05	NAIL
	−0.29	0.16	0.0	
HOLLOW	−0.29	0.22	0.08	PIN
	−0.29	0.16	0.0	
SCRIBBLE	−0.22	0.21	0.05	PEN
	−0.29	0.16	0.0	
SCRAWL	−0.24	0.24	0.02	CRAYON
	−0.29	0.16	0.0	

later used to ground the meaning of the word USE, while the remaining ones were non-iterative (Table 1b) and employed for the grounding of MAKE. The positions for which the joint values were recorded are shown in Table 1.

The CHOP–CUT actions are defined as a movement of the robot arm along the x axis, the HIT–POUND actions are defined along z, while the DRAW–PAINT on the y axis. The recorded joint values, before being sent to the neural network model, are normalized in the interval [0, 1] by using the following formula:

$$normalize(j_i) = \frac{j_i - J_{min}}{J_{max} - J_{min}} \quad (4)$$

where J_{min} and J_{max} represent the minimum and maximum values that the joint j_i to be normalized can assume. According to the joint values received in input, the neural network model computes the new joint values to be sent to the iCub robot; before being sent to the robot, the new joint values are de-normalized in the original interval, according to the following formula:

$$denormalize(j_i) = J_{min} + norm(j_i) \times (J_{max} - J_{min}) \quad (5)$$

Concerning the use of the appropriate grip (e.g. precision vs power grasp, that in our experiment were preprogrammed) it was selected depending on the dimension of the tool employed during the task. Tools of big dimensions required a power grasp, while for small tools a precision grasp was used.

4.2.3 Visual data set

The visual representations of features extracted from the objects used to perform actions are shown in Fig. 3a, b. Objects features are represented as a 4×4 binary matrix (M_{ij}) in which each element can assume either value 0 or 1. The features extracted from the objects presented to the robot are dimension, color and shape. The first element of the matrix (M_{11}) is related to the dimension of the object. The second, third and forth element of the matrix (M_{12}, M_{13}, M_{14}) encode the color of the object, while the remaining twelve elements $(M_{21}, M_{22}, M_{23}, M_{24}, M_{31}, M_{32}, M_{33}, M_{34}, M_{41}, M_{42}, M_{43}, M_{44})$ are related to the shape of the object. For example, the first binary matrix in (Fig. 3a) corresponds to the representation of a KNIFE with the following features: its dimension is small (M_{11} is 0), its color is red (M_{12}, M_{13}, M_{14} are 1, 0, 0) and its shape correspond to a predefined shape category (M_{21}, M_{22}, M_{23}, M_{24}, M_{31}, M_{32}, M_{33}, M_{34}, M_{41}, M_{42}, M_{43}, M_{44} are 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0).

We collected a training set that consists of 24 sequences including the motor, perceptual and linguistic inputs.

5 Training of the model

Studies conducted on developmental psychology and neurophysiology have revealed that perception and motor learning are pre-linguistic (Jeannerod 1997). That is, children acquire some motor behaviors and the capability to perceive objects before they learn to name them. In our experiments the iCub robot first develops some basic perceptual and motor skills necessary for initiating the interaction with the environment; hence, the robot can then use such knowledge to ground language. In particular, the robot is trained to recognize simple objects (e.g. KNIFE) and learn some higher-order behaviors (e.g. CUT). Following the approach used in our previous work (Cangelosi and Riga 2006; Stramandinoli et al. 2012), higher-order behaviors (e.g. CUT) are built based on the combination of basic motor primitives (e.g. PUSH and PULL). After the robot has acquired such simple visual and motor skills, it can use them for interacting in its environment

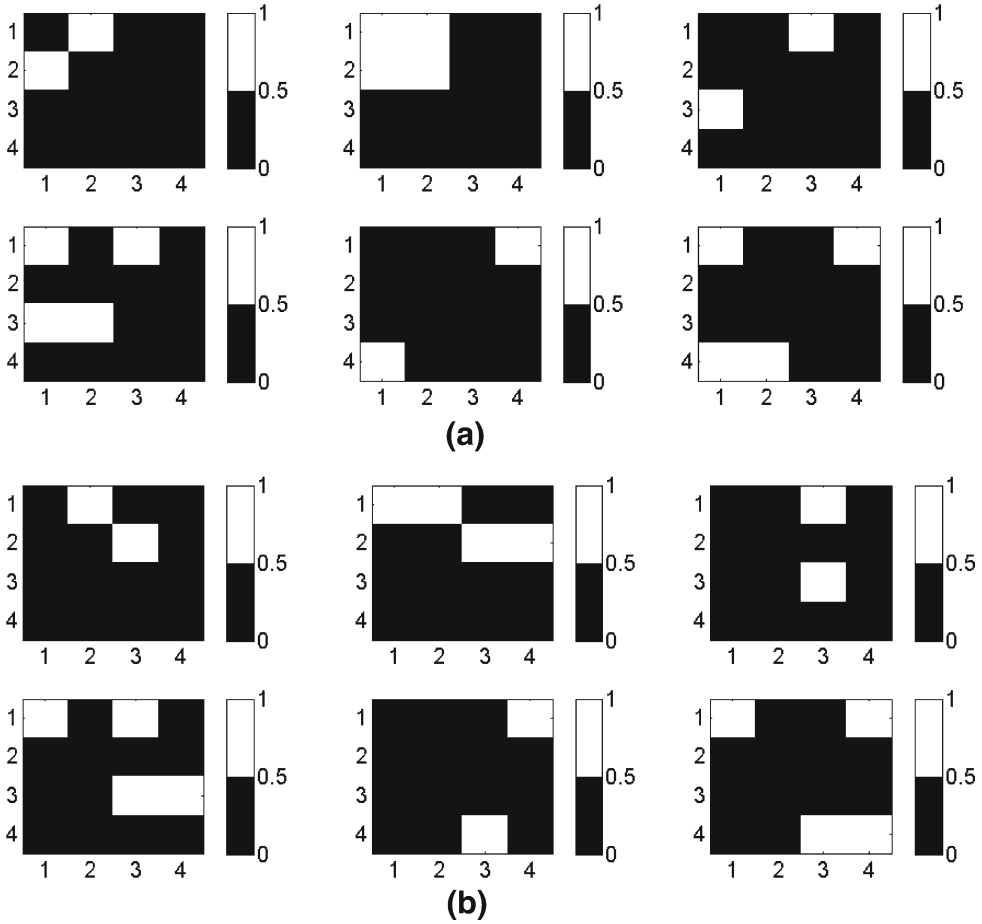

Fig. 3 Binary matrices representing the feature of the objects used to perform: iterative actions (**a**), non-iterative actions (**b**)

according to the received linguistic description. In Sect. 5.1 we describe the implemented training strategy.

5.1 Training stages

Taking inspiration by the aforementioned findings, the training of the neural network model was organized in three incremental stages:

1. *Pre-linguistic training*: The robot is trained to recognize a set of objects (e.g. KNIFE, HAMMER, BRUSH, etc.) and learn object-related actions (e.g. CUT, HIT, PAINT, etc.). Actions are both iterative and non-iterative and they are obtained by combining motor primitives; for example, the action primitive CUT is obtained by performing the motor primitives PUSH and PULL iteratively. During this stage the robot learns to recognize objects and perform actions independently from each others. That is, we do not train the robot to CUT [with] KNIFE but to learn how a KNIFE looks like, and how to perform the CUT action (independently from the usage of a specific tool). Table 1 contains the full list of objects and actions used for the training of the robot. The neural network model receives the proprioceptive input and the visual features of object's. The model outputs the next joint state and the representation of object's features.
2. *Linguistic-perceptual training*: This is the first stage of language acquisition. The model is trained to name actions and objects (two-words sentences consisting of a verb followed by a noun e.g. CUT [with] KNIFE); these words are directly grounded in perception and motor experience. The model, which was previously trained to perform action primitives and recognize object's features, during this stage receives in inputs the labels to be associated to actions and object's features. Given that in this stage the robot has to translate a linguistic commands (as CUT [with] KNIFE) into a behavior, it performs the action by using the appropriate tool.
3. *Linguistic-abstract training*: Abstract action words (i.e. USE, MAKE) are grounded by combining and recalling the perceptual and motor knowledge previously linked to basic words (i.e. Linguistic-perceptual training). To derive the meaning of abstract action words the robot, guided by linguistic instructions, organizes the knowledge directly grounded in perception and motor

knowledge. The model receives linguistic inputs related to abstract action words and outputs the corresponding behavioral patterns (i.e. next joint state). This phase of the training represents the abstract stage of language acquisition when new concepts are formed by combining the meaning of terms acquired during the previous stage of the training. Novel lexical terms can be continually acquired throughout the course of the robot's development through new sensorimotor interactions with the environment to which correspond new linguistic descriptions. Given that in this stage the robot has to translate a linguistic commands (as USE [a] KNIFE) into a behavior, the robot performs the action by using the appropriate tool.

At the end of the training, semantic meanings are formed via lexicon organization that recalls the perceptual knowledge and motor sequences in which the lexicon is grounded. In particular, the successful training of the model enables the robot to ground the meaning of words like USE and MAKE in the perceptual (e.g. features extracted from KNIFE, HAMMER, BRUSH, etc.) and motor experience (e.g. actions like CUT, HIT, DRAW) previously grounded. In the proposed hierarchical organization of lexical categories, words like KNIFE, HAMMER, CUT, HIT, etc., representing basic words, are directly grounded in perceptual and motor experience through a one-to-one mapping. Words like USE and MAKE, referring to different events and situations, are characterized by a one-to-many mapping; that is, a single linguistic label (e.g. USE) is associated to different basic words (KNIFE, PENCIL) (Borghi et al. 2011). We propose the hierarchical organization of concepts created by the model as a general and useful mechanism for the acquisition of abstract action words.

5.2 Learning algorithm

The aim of the training of the neural network model (Algorithm 5.1) is to ground the meaning of abstract action words in sensorimotor experience. In response to linguistic instructions the model has to generate the appropriate behavior and to recall the the representation of object's features. We define a function for evaluating the performance of the model in an offline mode; for such evaluation we select the mean square error (MSE). For the tuning of the neural network parameters we used the back-propagation algorithm. By finding the optimal values of the network weights that minimize the difference between the target and the actual output sequences, through the back-propagation algorithm the network learned the mapping between input and output values that permitted to perform the desired tasks. In the proposed study, the back-propagation algorithm was not used for mimicking the learning process of biological neural systems (Yamashita and Tani 2008), but rather as a general learning rule. Similar results could be obtained using other biologically more plausible learning algorithm (see Edelman 2015 for a proposal to reconsider some common assumptions made in the modelling of the brain and behavior).

Algorithm 5.1: BACK-PR. LEARNING($Data$)

```
Load net, params, data set
if simulation mode is training
  then Randomize initial weights [-0.1, 0.1]
for i ← 0 to maxCycles
do { Reset delta accumulation
     for p ← 0 to patternSetSequenceSize
       do { Reset inputs to 0
            Initialize state units to 0.5
            Learn the I/O mapping
     Update network's weights
     Compute MSE
     if MSE≤ threshold
       then Terminate algorithm
return (Output Values)
```

The maximum number of iterations of the learning algorithm was set to 10, 000. In order to avoid over-training of the network, the training was terminated as soon as the error reached the threshold value of 0.001 (stopping criterion). Indeed, the back-propagation algorithm as a possible stopping criteria includes that the total error of the network falls below a predefined threshold value or that a certain number of epochs are completed; in this work a combination of the two (i.e. whichever of the two occurs first) is used. The threshold value for the error was selected by training several networks and measuring the performance of each of them. The activation function of neurons in the hidden and output layers is a logistic function defined in the interval [0, 1]; the logistic function introduces non-linearity in the training and improves the convergence of the algorithm. The network's initial weights were drawn randomly from a uniform distribution defined in the interval $[-0.1, 0.1]$.

The training of the neural network model was implemented in batch mode according to which all the inputs in the training set are sent to the network before the weights are updated. For our work, the batch training has observed to be significantly faster and to produce smaller errors than the incremental training. Through the batch back-propagation, weight updates were summed over the presentation of the whole training sequences and subsequently the accumulated weight updates were performed. During each iteration of the algorithm, the accumulation of the variation of the weights was reset to zero; furthermore, for each pattern set the inputs were set to zero and the state units initialized to 0.5. Hence,

the new weight updates for the whole pattern set were computed for a certain number of epochs or until the stopping criterion was met (Algorithm 5.1).

Given the linguistic, proprioceptive and visual inputs, the model learns to predict the next joint state and produces the visual representations of objects. Carrying out several simulations, it has been possible to find the network's parameters that minimize the expected training and test error and hence to find the neural network model that allows the robot to properly perform the desired task.

6 Experimental results

Before presenting the results obtained on the neural network model described in Sect. 4, the evaluation settings are described.

6.1 Evaluation settings

The model was trained through back-propagation to learn the associations between words, and motor sequences and visual representations of objects; the training was performed for 25 random seeds. As already described in Sect. 5, the implemented training strategy consisted of three incremental stages, each of which corresponded to training the model in response to different configurations of the input signals. At the end of the second and third stage of the training, in order to understand how the model responded to the variation of the stimuli in input and further investigate how internal representations of objects are related to action representations, the performance of the model was evaluated in response to an *incompatible condition* test. We analysed the response of the model in case of inconsistency between the linguistic and visual inputs; objects and actions that the robot had previously learned to name, were referred using incompatible linguistic labels. In particular, two different incompatible condition tests were performed:

- *Incompatible noun condition*: to analyse the response of the model when the name of the object is incompatible with the object seen by the robot.
- *Incompatible verb condition*: to analyse the response of the model when the name of the action is incompatible with the behaviour that the robot usually performs with the presented object.

The linguistic input provided to the robot was either inconsistent with the objects perceived or with the actions typically associated to the presented objects. As such, this test was used to verify how the robot responded when the received linguistic command was in contrast with the perceived context.

In Sects. 6.2, 6.3 and 6.4 we present the results achieved during the three incremental stages of the training presented in Sect. 5.

6.2 Phase I: Pre-linguistic training

The first training stage of the model aimed at endowing the robot with basic perceptual and motor skills necessary for grounding higher-order concepts. The robot learned to classify objects according to their visual properties and to perform some predefined motor behaviors. In particular, the model was trained in a supervised manner to recognize 12 objects and perform 12 actions obtained by combining low-level motor primitives. The training was performed by activating the visual and proprioceptive inputs only, while the linguistic ones were silent. The training was successfully completed, and objects and actions were correctly categorized. The success of this training stage permitted the acquisition of the basic perceptual and motor knowledge necessary in the next stages of the training for the grounding of language.

6.3 Phase II: Linguistic-perceptual training

The second stage of the training enabled the robot to acquire linguistic capabilities through the direct naming of objects and actions. Connections between the motor/perceptual inputs and the linguistic labels were created.

6.3.1 Sensorimotor mapping

In order to have a quantitative measure of the similarity between the output and target joint values over time, the Dynamic Time Warping (DTW) was performed (Sakoe and Chiba 1978). In particular, we computed the DTW for the actual joint values produced by the model, and the target joint values used for the training of the robot. Results are presented in the gray-maps in (Fig. 4). Each row of the gray-map represents the actual joint values computed by the model, while columns represent the target joint values.

By displaying the results of the DTW in the gray-map layout in (Fig. 4), it is easier to visualise the capability of the model to categorize the proprioceptive inputs and analyse the performance of the robot in executing the desired behaviour. In particular, from Fig. 4a it is possible to observe that five out of the six iterative actions (i.e. CHOP, CUT, HIT, POUND, DRAW) have the lowest DTW values (corresponding to cell of darker gray in the map) when compared to their corresponding target values (cells on the main diagonal). For the PAINT action, the lowest DTW value is obtained when the output joint values are compared against the target joint values of the CUT action; this means that the robot, when asked to PAINT, it performs an action that in terms of joint values

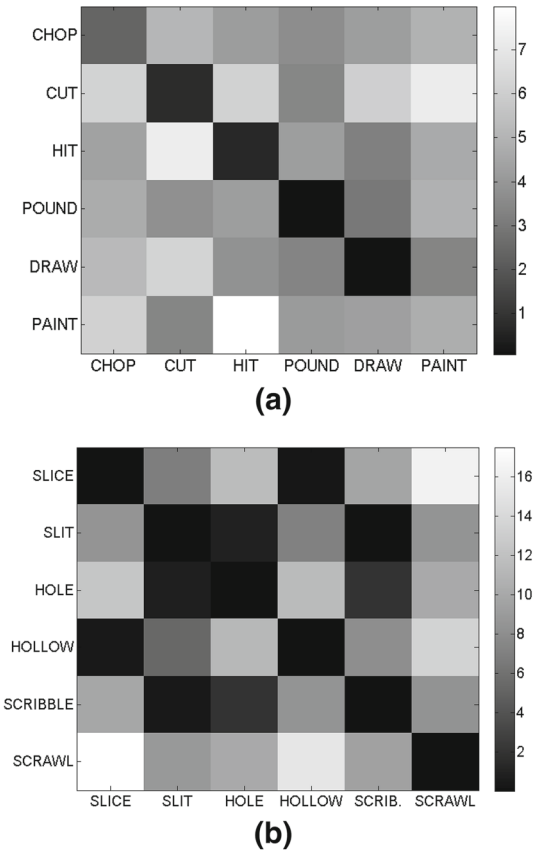

Fig. 4 Gray-map showing the DTW performed on joint values: iterative actions (**a**), non-iterative actions (**b**)

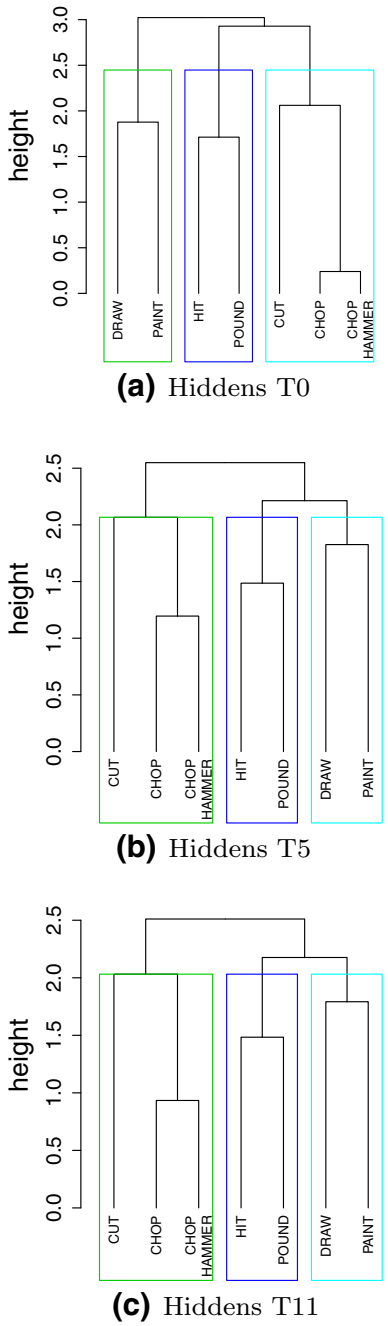

Fig. 5 Incompatible *noun* condition (e.g."**CHOP** [with] **KNIFE**" *became* "**CHOP** [with] **HAMMER**"). Results of the hierarchical clustering of hidden units at the time steps $T = 0$ (**a**), $T = 5$ (**b**), and $T = 11$ (**c**)

is closer to the CUT than the PAINT action. From Fig. 4b it is possible to observe that all the six non-iterative actions were very well performed and classified. Given the similarity among the six non-iterative actions, the DTW has low values in correspondence of more than one target; nevertheless, in this case the lowest DTW is registered on the main diagonal of the gray-map (Fig. 4b).

The performance of the robot in terms of action execution for the non-iterative actions is better than the iterative ones. However, the mapping of the joint values associated to the non-iterative actions was easier than learning the mapping of the joint values associated to the iterative ones, which required to repetitively alternate the values of the robot's encoders from the home to the target values.

6.3.2 Incompatible condition test

Activation values of hidden units recorded during the incompatible condition test were analysed. In particular, in order to compare the hidden activation values recorded at each time step during the compatible and incompatible conditions, a temporal hierarchical cluster analysis has been performed. As a measure of dissimilarity between pairs of observations, the Euclidean distance has been used. Due to lack of space, in this paper we show only results of the hierarchical cluster analysis for the timesteps 0, 5, 11. However, during the other timesteps the obtained dendrograms are either equal to the one for timestep 0 or 5–11.

Incompatible noun condition test The results of the hierarchical clustering of hidden values at the time steps $T = 0$, $T = 5$ and $T = 11$ are presented in (Fig. 5), where

dendrograms compare the activation values recorded during the compatible condition "**CHOP** [with] **KNIFE**" to the activation values recorded during the incompatible condition "**CHOP** [with] **HAMMER**". The incompatibility is related to the KNIFE/HAMMER nouns. Despite the robot seeing a KNIFE, the word HAMMER is used to refer to the object. The dendrograms in (Fig. 5) show that the observations are organized in three main clusters that pair the inputs related to the six iterative actions. The hidden values related to the incompatible condition "**CHOP** [with] **HAMMER**" are clustered together with CHOP. This means that the activation values of hidden units during this incompatible condition are similar to the activation values of hidden units recorded during the compatible condition "**CHOP** [with] **KNIFE**".

Incompatible verb condition test The results of the hierarchical clustering of hidden values at the time steps $T = 0$, $T = 5$ and $T = 11$ are presented in (Fig. 6); such dendrograms compare the activation values recorded during the compatible condition "**CHOP** [with] **KNIFE**" to the activation values recorded during the incompatible condition "**DRAW** [with] **KNIFE**". In this test the inconsistency is related to the substitution of the verb CHOP with DRAW. Despite in front of the robot there is a KNIFE, the verb DRAW is used to refer to the action to be performed with the presented object. The dendrograms in (Fig. 6) show that the observations are organized in three main clusters that pair the inputs related to the six iterative actions. The activation values related to the incompatible condition "**DRAW** [with] **KNIFE**" are clustered together with CHOP. This means that the activation values of hidden units during this incompatible condition test are similar to those recorded during the compatible condition. The incompatible condition tests seem to suggest that, in case of inconsistency, the perceptual input is stronger than the linguistic one and it triggers the behaviour expected to be performed with a specific object. The results of this test can be helpful in understanding the mechanisms underlying positive, as well as, negative compatibility effects observed in behavioural experiments (Borghi et al. 2004; Tucker and Ellis 2004).

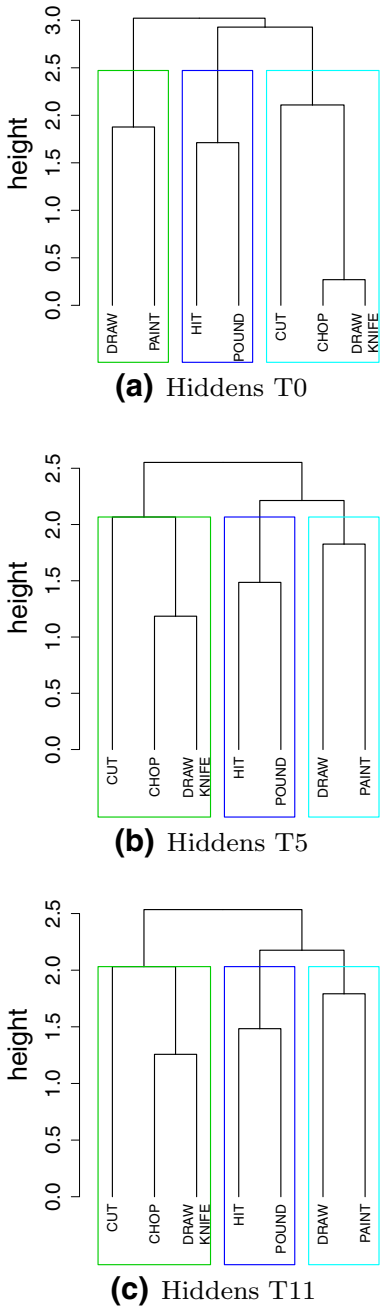

Fig. 6 Incompatible *verb* condition (e.g. "**CHOP** [with] **KNIFE**" *became* "**DRAW** [with] **KNIFE**"). Results of the hierarchical clustering of hidden units at the time steps $T = 0$ (**a**), $T = 5$ (**b**), and $T = 11$ (**c**)

6.4 Phase III: Linguistic-abstract training

The last stage of the training has enabled the iCub to learn abstract action words and acquire higher-order categories. New concepts were formed by combining the lexical terms acquired during the previous stage of the training. Since such lexical terms are directly connected to perceptual and motor experience, they recall the previously grounded perceptual and motor knowledge (multi-modal symbols).

6.4.1 Sensorimotor mapping

The similarity between the output and target joint values over time has been calculated by performing the DTW. The output joint values, recorded after each action, were compared to the corresponding target values (Fig. 7). For both iterative and non-iterative actions it is possible to observe that the lowest DTW is obtained when the actual output joint val-

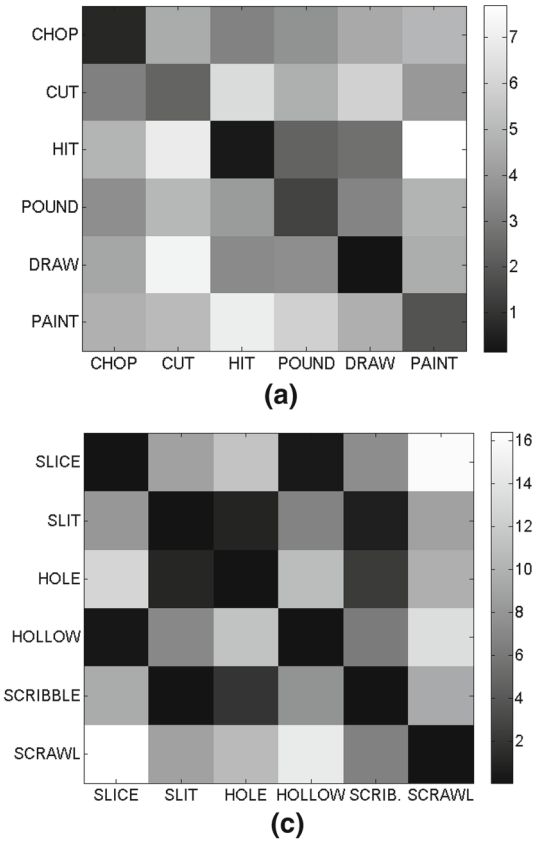

Fig. 7 Gray-map showing the DTW performed on joint values: iterative actions (**a**), non-iterative actions (**b**)

ues are compared to their corresponding targets (Fig. 7a, b). The performance of the robot in executing action primitives improves after the third stage of the training.

6.4.2 Incompatible noun condition test

We analysed the response of the model in case of inconsistency between the linguistic and visual inputs. In particular, the incompatible noun condition was tested to analyse the response of the model when the name of the object is incompatible with the object perceived by the robot (e.g. "**USE** [a] **KNIFE**" *became* "**USE** [a] **HAMMER**"). Activation values of hidden units recorded during the compatible and incompatible conditions were analysed by performing the temporal hierarchical cluster analysis. Figure 8 shows the results of the hierarchical clustering of hidden units at the time steps $T = 0$, $T = 5$ and $T = 11$; the dendrograms compare the hidden activation values recorded during the compatible condition "**USE** [a] **KNIFE**" to the hidden activation values recorded during the incompatible condition "**USE** [a] **HAMMER**". In this case, the incompatibility is related to the KNIFE/HAMMER nouns. Despite the robot sees a KNIFE, the word HAMMER is used to refer to the object placed in

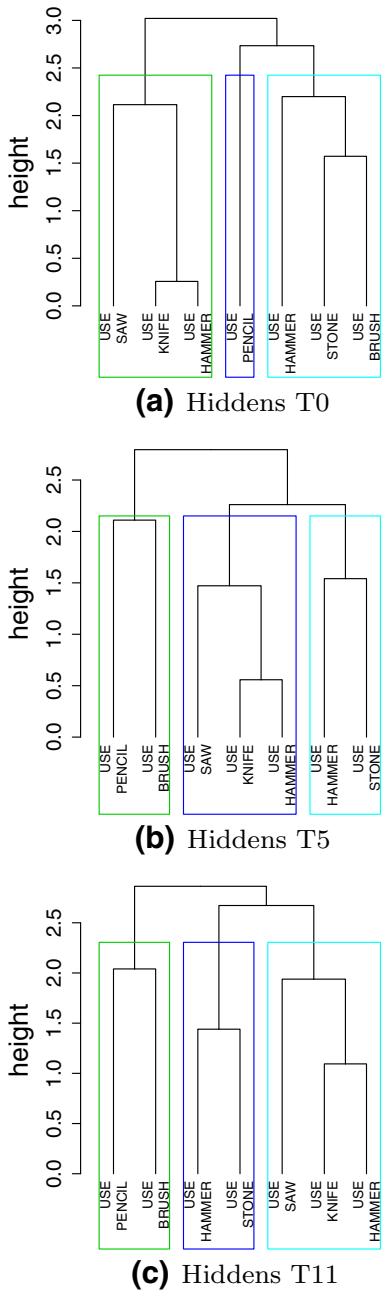

Fig. 8 Incompatible noun condition (e.g. "**USE** [a] **KNIFE**" *became* "**USE** [a] **HAMMER**")). Results of the hierarchical clustering of hidden units at the time steps $T = 0$ (**a**), $T = 5$ (**b**), and $T = 11$ (**c**)

front of the robot. The hidden values related to the incompatible condition "**USE** [a] **HAMMER**" are clustered together with "**USE** [a] **KNIFE**". This means that the activation values of hidden units during this incompatible condition test are very close to the activation values of hidden units recorded during the compatible condition.

The results obtained in the incompatible noun condition test has confirmed that in the case of inconsistency between the perceptual and linguistic input, the robot executes the

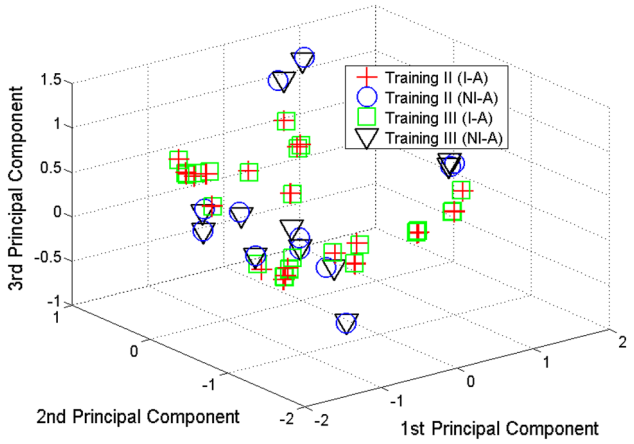

Fig. 9 Hidden units activation values in the space of the three principal components. Data displayed in four groups: Training II Iterative-Actions, Training II Non-Iterative Actions, Training III Iterative-Actions, and Training III Non-Iterative-Actions

actions elicited by the seen objects. This suggests that the proper naming of objects and actions supports action categorization and that seeing objects automatically elicits the representations of their affordances (i.e. all the motor acts that can be executed on particular objects to obtain a desired effect).

6.5 Representations of abstract action words

After all the stages of the training were successfully completed, to better understand the internal dynamics of the model, the activation values of hidden units were analysed by performing the Principal Component Analysis (PCA). Observations are displayed in four groups representing respectively the activation values in the space of the first three principal components for the Training II and III for Iterative-Actions, and the Training II and III for Non-Iterative-Actions (Fig. 9). The observations related to the iterative actions recorded during the second and third stage of the training almost fully overlap (e.g. data displayed by red and green markers). The same consideration can be done for non-iterative actions (e.g. data displayed by blue and black markers). This confirms that hidden units during the second and third stage of the training follow a very similar activation pattern. These results suggest that the acquisition of concepts related to abstract action words (e.g. USE and MAKE) requires the reactivation of similar internal representations activated during the acquisition of the basic concepts that are hierarchically organized to ground a particular abstract action word. This seems to suggest that even the semantic/conceptual representation of abstract action words requires reusing motor and perceptual representational capabilities (Barsalou 1999).

7 Conclusions

In this work we proposed a model for the acquisition of abstract action words grounded in the in perceptual and motor knowledge of a humanoid robot. Although the proposed experimental setup is limited, given the exemplification made in the representation of the multi-modal inputs, it suggests a general mechanism for grounding abstract action words through the combination of perceptual knowledge and simple motor primitives in humanoid robots. The implemented architecture is based on partially recurrent neural networks (Jordan 1986), which enabled the modelling of the mechanisms underlying motor and linguistic sequence processing. The training of the model was incremental and consisted of three stages that permitted to acquire perceptual and motor knowledge first, to learn words directly grounded in perceptual and motor knowledge subsequently, and to ground abstract action words through the hierarchical organization of the words directly linked to perceptual and motor knowledge at the end.

Experimental results have shown that the robot was able to perform the behaviour triggered by the linguistic input and the perceived object; the joint values produced by the robot were not identical to the values of the ones used for the training, but the difference was still acceptable to reproduce the requested behavior. The presence of clusters in the hidden units of the model suggested the formation of concepts from the multi-modal data received in input by the network. Results obtained in the incompatible condition tests showed that in case of inconsistency between the perceptual and linguistic inputs, the robot executed the actions elicited by the seen object.

Directions for future research include the grounding of language in tool affordances through statistical inference. Despite being clear that language needs to be grounded in sensorimotor experience, it is also necessary to go beyond simple sensorimotor grounding (Thill et al. 2014). To this end, statistical inference will be adopted in grounded theories of meaning. Embodied theories of meanings in a probabilistic framework can lead to "hybrid models" in which some concepts are directly grounded in a robot's sensorimotor experience while, for other concepts, statistical inference will permit to go beyond the available data and acquire new concepts.

Acknowledgements This research was supported by the Marie Curie Initial Training Network RobotDoC (235065) and the Marie Curie Intra European Fellowship RoboTAsk (624424) within the 7th European Community Framework Programme, and the EPSRC BABEL project.

References

Arbib, M. A., Érdi, P., et al. (1998). *Neural organization: Structure, function, and dynamics*. New York: MIT Press.

Barsalou, L.W., Santos, A., Simmons, W.K., Wilson, C.D., De Vega, M., Glenberg, A., Graesser, A. (2008). Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, pp. 245–283.

Barsalou, W. L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(04), 577–660.

Barsalou, W. L. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Barsalou, W. L. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1281–1289.

Barsalou, W. L., Simmons, W. K., Barbey, A. K., & Wilson, D. C. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84–91.

Borghi, A.M., & Binkofski, F. (2014). The wat proposal and the role of language. In *Words as social tools: An embodied view on abstract concepts* (pp. 19–37). New York: Springer.

Borghi, A. M., Capirci, O., Gianfreda, G., & Volterra, V. (2014). The body and the fading away of abstract concepts and words: A sign language analysis. *Frontiers in Psychology*, *5*, 811.

Borghi, M. A., Glenberg, M. A., & Kaschak, P. M. (2004). Putting words in perspective. *Memory & Cognition*, *32*(06), 863–873.

Borghi, M.A., Flumini, A., Cimatti, F., Marocco, D., Scorolli, C. (2011). Manipulating objects and telling words: A study on concrete and abstract words acquisition. *Frontiers in Psychology*

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., & Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study. *Cognitive Brain Research*, *24*(3), 355–363.

Cangelosi, A., & Parisi, D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, *89*(2), 401–408.

Cangelosi, A., & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive science*, *30*(4), 673–689.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., et al. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, *2*(3), 167–195.

DeWolf, T., & Eliasmith, C. (2011). The neural optimal control hierarchy for motor control. *Journal of Neural Engineering*, *8*(6), 065,009.

Dominey, F. P., Mallet, A., & Yoshida, E. (2009). Real-time spoken-language programming for cooperative interaction with a humanoid apprentice. *International Journal of Humanoid Robotics*, *6*(02), 147–171.

Dove, G. (2011). On the need for embodied and dis-embodied cognition. *Embodied and grounded cognition*, p. 129.

Edelman, S. (2015). The minority report: Some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–26.

Elman, L. J. (1990). Finding structure in time. *Cognitive science*, *14*(02), 179–211.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3–4), 455–479.

Glenberg, M. A., & Kaschak, P. M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558–565.

Glenberg, M. A., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., & Buccino, G. (2008). Processing abstract language modulates motor system activity. *The Quarterly Journal of Experimental Psychology*, *61*(6), 905–919.

Grafton, S. T., & Hamilton, A Fd C. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, *26*(4), 590–616.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307.

Heit, E., & Barsalou, L. W. (1996). The instantiation principle in natural categories. *Memory*, *4*(4), 413–451.

Jeannerod, M. (1997). *The cognitive neuroscience of action*. Oxford: Blackwell.

Jordan, I.M. (1986). Serial order: A parallel distributed processing approach. Institute for Cognitive Science Report 8604, University of California, San Diego.

Kalkan, S., Dag, N., Yürüten, O., Borghi, M.A., Sahin, E. (2013). Verb concepts from affordances. *Interaction Studies Journal*.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by, vol. 111. London: Chicago.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

McGhee-Bidlack, B., et al. (1991). The development of noun definitions: A metalinguistic analysis. *Journal of Child Language*, *18*(02), 417–434.

Metta, G., Fitzpatrick, P., & Natale, L. (2006). Yarp: Yet another robot platform. *International Journal on Advanced Robotics Systems*, *3*(01), 43–48.

Metta, G., Sandini, G., Vernon, D., Natale, L., Nori, F. (2008). The icub humanoid robot: An open platform for research in embodied cognition. In: Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems (pp. 50–56). London: ACM.

Murphy, G. L., & Wisniewski, E. J. (1989). Categorizing objects in isolation and in scenes: What a superordinate is good for. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 572.

Mussa-Ivaldi, A. F., & Bizzi, E. (2000). Motor learning through the combination of primitives. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, *355*(1404), 1755–1769.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 211.

Oztop, E., & Arbib, A. M. (2002). Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics*, *87*(02), 116–140.

Pattacini, U., Nori, F., Natale, L., Metta, G., Sandini, G. (2010). An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1668–1674). New York: IEEE.

Perani, D., Cappa, F. S., Schnur, T., Tettamanti, M., Collina, S., Rosa, M. M., et al. (1999). The neural correlates of verb and noun processing. *Brain*, *122*(12), 2337–2344.

Pinker, S. (1994). *The language instinct: How the mind creates languages*. New York: HarperCollinsRickford.

Pulvermüller, F., Härle, M., & Hummel, F. (2001). Walking or talking?: Behavioral and neurophysiological correlates of action verb processing* 1. *Brain and Language*, *78*(2), 143–168.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*(2), 131–141.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.
Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *01*, 43–49.
Scorolli, C., & Borghi, M. A. (2007). Sentence comprehension and action: Effector specific modulation of the motor system. *Brain Research*, *1130*, 119–124.
Stramandinoli, F. (2014). Towards the grounding of abstract categories in cognitive robots. PhD thesis.
Stramandinoli, F., Cangelosi, A., Marocco, D. (2010). Towards the grounding of abstract words: A neural network model for cognitive robots. In *Proceedings of IJCNN-2011 international joint conference on neural networks*
Stramandinoli, F., Marocco, D., & Cangelosi, A. (2012). The grounding of higher order concepts in action and language: A cognitive robotics model. *Neural Networks*, *32*, 165–173.
Sugita, Y., & Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior*, *13*(01), 33.
Tettamanti, M., Buccino, G., Saccuman, C. M., Gallese, V., Danna, M., Scifo, P., et al. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*(2), 273–281.
Thill, S., Pad, S., & Ziemke, T. (2014). On the importance of a rich embodiment in the grounding of concepts: Perspectives from embodied cognitive science and computational linguistics. *Topics in Cognitive Science*, *6*(3), 545–558.
Tucker, M., & Ellis, R. (2004). Action priming by briefly presented objects. *Acta Psychologica*, *116*(2), 185–203.
Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science*, *29*(5), 719–736.
Wiemer-Hastings, K., Krug, J., Xu, X. (2001). Imagery, context availability, contextual constraint, and abstractness. In *Proceedings of the 23rd annual conference of the cognitive science society*, pp. 1134–1139.
Yamashita, Y., & Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. *PLoS Computational Biology*, *4*(11).
Yürüten, O., Uyanık, F.K., Çalışkan, Y., Bozcuoğlu, K.A., Şahin, E., Kalkan, S. (2012). Learning adjectives and nouns from affordances on the icub humanoid robot. In *From animals to animals*, vol. 12 (pp. 330–340). New York: Springer.

Francesca Stramandinoli has a MEng in Automation Engineering from the University of Calabria (Italy). In 2009 she worked as Research Assistant at the Department of Mathematics of the University of Calabria, where she dealt with the design of CNN-based algorithms for medical images segmentation and the numerical integration of non-linear partial differential equations. From 2010 up to 2013 she worked as Marie Curie Early Stage Researcher at Plymouth University (United Kingdom) within the RobotDoC European Initial Training Network. Her research topic revolved around the grounding of language in humanoid robots with particular attention to abstract words. Stramandinoli received her Ph.D. in Computing from Plymouth University (United Kingdom) in 2014. After getting her Ph.D. she worked for one year as Research Engineer at the ALES S.r.l. (Advanced Laboratory on Embedded Systems) and UTSCE (United Technologies Systems & Controls Engineering) in Rome (Italy) where she contributed to the development of technologies that support the application of simulation-based contract verification to System of systems. Currently she is a Marie Curie Experienced Researcher within the RoboTAsk Intra-European Fellowship at the Italian Institute of Technology (Italy); she works on the learning of action words and tool affordances in the iCub humanoid robot.

Davide Marocco received his PhD in Artificial Intelligence at the University of Calabria, Italy, in 2004. He is currently Associate Professor (Reader) of Cognitive Robotics and Intelligent Systems at the University of Plymouth, UK and the coordinator of the local CUDA Teaching Centre. His research interests are mainly focused on evolutionary robotics models of behavior and evolution of communication and language. However, given his interests on CUDA and GPU parallel programming, he has side collaboration on the application of CUDA with cellular automata and genetic algorithms, recently applied to lava flows discrete models.

Angelo Cangelosi is Professor of Artificial Intelligence and Cognition and the Director of the Centre for Robotics and Neural Systems at Plymouth University (UK). Cangelosi's main research expertise is on language grounding and embodiment in humanoid robots, developmental robotics, human-robot interaction, and on the application of neuromorphic systems for robot learning. He currently is the coordinator of the UK EPSRC project "BABEL: Bio-inspired Architecture for Brain Embodied Language" (2012–2016). He also is Principal investigator for the ongoing projects "THRIVE" (US Air Force Office of Science and Research, 2014–2018), the FP7 projects POETICON++ and ROBOT-ERA, and the Marie Curie projects SECURE, ORATOR and DECORO. In 2012–2013 he was Chair of the IEEE Technical Committee on Autonomous Mental Development. Cangelosi is Editor-in-Chief of the IEEE Transactions on Autonomous Mental Development, and also is Editor of the journal Interaction Studies. His latest book "Developmental Robotics: From Babies to Robots" (MIT Press; co-authored with Matt Schlesinger) has just been released, as of January 2015.